

Data Retention & Sub-processor Statement

CIATA DeepResearch

Version: 2.9 **Date:** 2026-10-08 **Controller:** CIATA GmbH, Arnimallee 7, 14195 Berlin, Germany · HRB 266084 B · Amtsgericht Charlottenburg · MD: Prof. Dr. Tim Landgraf · privacy@ciata.io

Revision 2.9 (2026-10-08): Staan's DPA is on file (Staan DPA, signed_dpas/staan/). It does not support the zero-retention statement: Staan keeps the queries for 25 months on OVH servers in France (DPA Appendix 1) and treats them as not personal data (DPA §2). Staan is no longer listed as a zero-retention step; §1, §3 and §5.2 updated. The provider is European Perspective SAS, Paris, not "European Search Perspective GmbH". Staan's FAQ says an unnamed US fallback can supplement results, so some queries may leave the EU; the DPA does not list it. CIATA has asked Staan to switch it off for its account (issue #194).

Revision 2.8 (2026-10-04): Server and access logs: each deploy now saves the container log before replacing the container, as a compressed archive on the same server deleted 90 days after the container started (issue #161). The lifecycle table stated "~90 days" without a mechanism; no Docker log rotation is configured, and until now a log lived only until the next deploy. Research progress messages (questions, sub-questions, search queries, source titles) are no longer written to the server log (issue #164): they go to a log file inside the run's own directory, which is deleted with the run after 30 days. The server log keeps operational events and warnings, with e-mail addresses hashed or redacted and web addresses cut to their host name.

Revision 2.7 (2026-09-27): Consistency Checker added (issue #172): runs in parallel with the Paragraph Checker on qwen3.8-27b with reasoning effort medium; no new model, processor or data category; two additional LLM calls per report. §5.1 updated: Consistency Checker added to the qwen3.8-27b agent list.

Revision 2.6 (2026-09-27): Language Guard moves from qwen3.6-chat-35b to gemma4-chat-26b with reasoning disabled (issue #167). qwen3.6-chat-35b is no longer used in any agent; three models remain in use (gpt-oss-120b, qwen3.8-27b, gemma4-chat-26b). Failed repairs now remove leaked characters in code rather than keeping the line. §5.1 updated: model list for Language Guard changed; "four" changed to "three"; qwen3.6-chat-35b noted as used until 2026-09-27. Same data categories, same recipient (AKI.io GmbH, Frankfurt).

Revision 2.5 (2026-09-26): Transactional email moves from Brevo SAS to Scaleway SAS (issues #152, #159). Brevo remains in the code as a rollback option; after the switch no personal data is sent to Brevo. §5.7 updated: Scaleway SAS replaces Brevo SAS as the active email processor; Brevo is noted only as former processor, with its wind-down as an open item. The report-ready mail (issue #159, live since 2026-09-26), which carries the report title,

a short summary, the report link and the deletion date, is now sent via the email processor; it is on by default and can be disabled in account settings. The lifecycle table and §5.7 are updated to include the report-ready mail's data.

Revision 2.4 (2026-09-13): Model routing revised for run time (issue #135): the Relevance Gate moves from qwen3.6-chat-35b to gpt-oss-120b (same verdicts on 192 replayed results, 4.5x faster) and the Extractor from qwen3.6-chat-35b to gemma4-chat-26b with reasoning disabled (5x faster, more and better-grounded findings on 40 real pages). gemma4-chat-26b is a NEW model identifier — a fourth model now processes customer data (PII-filtered page excerpts) — so the DPA scope must name four models; qwen3.6-chat-35b remains in service for the Language Guard only. Same data categories, same recipient (AKI.io GmbH, Frankfurt).

Revision 2.3 (2026-09-12): Post-synthesis verification restructured (issue #133): the Fact Checker — a whole-report check plus a per-citation alignment check, both on gpt-oss-120b — is replaced by the Paragraph Checker, and the Report Patcher is given its own routing; both now run on qwen3.8-27b. Same data categories (the draft report and the run's findings), same recipient (AKI.io GmbH), no new model identifier: the three-model set and the DPA scope instruction are unchanged.

Revision 2.2 (2026-09-10): Synthesis model corrected from mistral4-119b to qwen3.8-27b, in service since 2026-08-23 (commit eb2a5b1); the Table Repair agent, added 2026-09-10, runs on that same model and introduces no new recipient. DPA scope instructions corrected in step. The model list in §5 named mistral4-119b from 2026-08-23 until this revision.

Revision 2.1 (2026-08-11): (1) Corrected page-content persistence: PII-filtered extractor input is written to state.json (raw/unfiltered text is never persisted); lifecycle table and §7 updated accordingly. (2) Fixed §9 audit commands — retention-clock command now reads created_at from job_meta.json (the governing timestamp); page-text command now inspects actual page_text fields. (3) §8 names created_at in job_meta.json as the retention clock. (4) §5.8 corrected to describe Plausible click-level conversion events. (5) Added run provenance metadata row to lifecycle table. (6) AKI.io model list unchanged (already corrected in v2.0). (7) Added §5.9 recording the planned — not yet enabled — OpenObserve operational telemetry, with its pre-enablement conditions. (8) Qualified references to the AKI.io and Staan zero-retention commitments: those DPAs are not yet executed.

Scope note — "Zero retention" is not a system-wide claim. The phrase zero data retention applies narrowly to **one specific processing step: AKI.io prompt processing** — LLM prompts and completions are not stored or used for training by AKI.io. This is the **intended contractual basis**; the Art. 28 data processing agreement giving it effect is **not yet executed** (see §5.1), so it should not be read as a confirmed contractual guarantee today.

Staan is **not** a zero-retention step: its DPA keeps search queries for 25 months (see §5.2).

The CIATA DeepResearch system as a whole uses **time-bounded retention**: research reports are retained for 30 days (configurable), payment records for 10 years (§ 147 AO), and other categories as set out in the lifecycle table in Section 3.

1. Executive Summary

CIATA DeepResearch implements a **time-bounded data retention model** with configurable TTL (default: 30 days). Research data is automatically purged after the retention period. Research prompts sent to AKI.io (LLM) are not retained by AKI.io beyond the transaction. Search queries sent to Staan are kept by Staan for 25 months (Staan DPA, Appendix 1). Payment processing involves additional sub-processors (Mollie, Billie, Qonto, Scaleway) that retain data for their own regulatory and contractual purposes beyond CIATA's retention schedule; these are described in Section 5.

2. Data Flow Architecture

```
sequenceDiagram
    participant U as User (Browser)
    participant API as CIATA API (Hetzner Cloud DE)
    participant Staan as Staan Search (EU)
    participant Web as Public Web
    participant LLM as AKI.io LLM (Frankfurt DE)
    participant Qonto as Qonto Invoicing (FR)
    participant Scaleway as Scaleway Email (FR)

    U->>API: Research question
    API->>API: Decompose question (LLM call)
    API->>Staan: Search queries + full_content=markdown
    Note over Staan: Staan server-side fetches target pages<br/>and returns full page
    text (markdown)
    Staan-->>API: URLs + snippets + full page text
    API->>Web: Direct page fetch (robots.txt checked first)
    Note over Web: Only when Staan content absent or insufficient
    Web-->>API: HTML content
    Note over API: PII filter applied to all page content
    API->>LLM: Extract findings (scrubbed content)
    LLM-->>API: Structured findings
    Note over LLM: No data retention by AKI.io
    API->>LLM: Synthesize report
    LLM-->>API: Report draft
    API->>U: Final report (Markdown, VSRMark-watermarked)
    Note over API: Run data auto-deleted after 30 days
    API->>Qonto: Invoice creation (name, email, billing address)
    API->>Scaleway: Purchase confirmation email, report-ready mail
```

3. Data Lifecycle

Stage	Data	Retention	Location
In transit	Research question, LLM prompts	Transient (request duration)	TLS-encrypted
LLM processing	Prompts + completions	Zero — AKI.io does not retain (intended contractual basis; DPA not yet executed)	AKI.io Frankfurt, DE
Search queries	Generated queries	25 months at Staan (Staan DPA, Appendix 1)	Staan, OVH France
Staan full_content	Page text fetched by Staan server-side; PII filter applied in-process before any disk write	Transient at Staan; PII-filtered extractor input written to <code>state.json</code> (see row below)	Staan EU → Hetzner Cloud DE
Page content (direct)	HTML text fetched by CIATA crawler; PII filter applied in-process before any disk write	Transient (unfiltered text is never written to disk); PII-filtered extractor input written to <code>state.json</code> (see row below)	Hetzner Cloud DE
Stored extractor input	PII-filtered page text (capped at 16,000 chars/source, <code>page_max_chars</code>); the exact text the LLM extractor received, retained for diagnostic use (zero-yield source analysis)	30 days, deleted with the run (<code>DATA_RETENTION_DAYS</code>)	<code>runs/{job_id}/state.json</code> , Hetzner Cloud DE, disk; readable only via <code>GET /v1/admin/reports/{job_id}/sources/{i}/text</code> , which requires the caller to hold the admin role AND own that specific run — no staff override exists
Extracted findings	Facts from pages (PII-filtered)	30 days (configurable, <code>DATA_RETENTION_DAYS</code>)	Hetzner Cloud DE, disk
Research report	Markdown report	30 days (configurable)	Hetzner Cloud DE, disk
State file	Full research state (includes stored extractor input and run provenance metadata — see rows above and below)	30 days (configurable)	Hetzner Cloud DE, disk
Account data	Email, password hash	Until account deletion	Hetzner Cloud DE, SQLite
Unverified accounts	Email + verification token	72 hours	Hetzner Cloud DE
JWT refresh tokens	Auth tokens	7 days	Hetzner Cloud DE

Stage	Data	Retention	Location
Server / access logs	IP, request metadata	At most 90 days: the container log until the next deploy, then a compressed archive (/var/log/deepresearch/, root-only) deleted 90 days after the container started (issue #161)	Hetzner Cloud DE
Payment data	Transaction references	10 years (§ 147 AO)	Mollie B.V. (NL)
Invoicing data	Name, email, billing address	Qonto retains for own accounting/tax purposes	Qonto SAS (FR)
B2B underwriting data	Billing address, VAT/ Handelsregister	Billie retains for own underwriting/risk purposes	Billie GmbH (DE)
Transactional email logs	Email address, name; purchase details (credits, pack label, price, new balance, Mollie payment reference, purchase date). Report-ready mail (live since 2026-09-26, issue #159): additionally report title, a short summary of the report (the first paragraph of its executive summary, up to approximately 500 characters), link to the report, and the report's deletion date. The report-ready mail can be turned off in account settings (Notifications).	⚠ TO CONFIRM: retention period of Scaleway TEM sending logs (per Scaleway's DPA)	Scaleway SAS, fr-par region, Paris, France. Former processor: Brevo SAS (FR), held logs until switch date (2026-09-26).
Run provenance metadata	schema_version, code_revision, agent_models, config_digest – operational metadata recording which version of the system produced the run; not personal data	30 days, deleted with the run	runs/{job_id}/state.json, Hetzner Cloud DE, disk
Analytics	Page-view counts, referrer statistics, and click-level conversion events (event name and URL only; no personal data, no form values, no identifiers)	CIATA-controlled, self-hosted	trc.ciata.io (Hetzner Cloud DE)

Note on Staan full_content: The default configuration (SEARCH_BACKEND=staan, STAA_N_FULL_CONTENT=true) causes Staan's servers to fetch the full text of each result page server-side before returning it to CIATA. Web page content therefore passes through Staan's infrastructure before CIATA's own PII filter is applied. CIATA's PII filter is then applied to the markdown received from Staan. See Section 4 for filter scope and limitations.

4. PII Filter Scope

The PII filter (`src/pii_filter.py`) is applied to all web content (whether received from Staan or fetched directly) before it is sent to the LLM for extraction. It covers:

- Email addresses
- International phone numbers (DACH-tuned patterns)
- German Steuerliche Identifikationsnummer (11 digits)
- German Sozialversicherungsnummer
- Swiss AHV/AVS numbers
- IBANs
- Luhn-validated credit card numbers
- IPv4 addresses
- German-format postal addresses (Straßenname + Hausnummer)

Out of scope (not currently filtered): personal names, non-DACH national identifiers (e.g. UK National Insurance, US SSN, FR/NL/PL national IDs), and non-German postal address formats.

5. Third-Party Sub-processors

5.1 AKI.io GmbH (LLM Provider)

- **Location:** Frankfurt, Germany
- **Data sent:** Research prompts containing PII-filtered page content and research questions. Three models are in use: gpt-oss-120b (reasoning agents and the relevance gate), qwen3.8-27b (report synthesis, table repair, paragraph check, consistency check and patch), gemma4-chat-26b (extraction and language guard, reasoning disabled). All three route through AKI.io infrastructure in Frankfurt. qwen3.6-chat-35b was used for the language guard until 2026-09-27 and is no longer active. minimax-m2.5-230b, named in earlier versions of this statement, is no longer used anywhere in the pipeline (removed 2026-07-26; AKI.io is retiring it).
- **Retention:** Zero — prompts and completions are not stored or used for training.
- **Legal basis:** Art. 6(1)(b) GDPR (contract performance); Art. 28 GDPR DPA (required — not yet executed); processing within EU.
- **DPA:** Required — not yet executed / not on file.

△ **TO CONFIRM:** Confirm AKI.io DPA clause reference that establishes (a) no prompt retention and (b) no use for model training, and confirm DPA execution date, for inclusion in the Art. 30 records of processing activities.

5.2 Staan / European Perspective SAS (Search Provider)

- **Location:** EU (Ecosia + Qwant index; European Perspective SAS)

- **Data sent:** Search queries derived from the user's research question. When `full_content=markdown` is active (the default), Staan also server-side fetches result pages and returns their full text to CIATA — web page content therefore passes through Staan's servers prior to CIATA's PII filter.
- **Retention:** Queries are kept for 25 months (DPA Appendix 1). The DPA does not mention the `full_content` page fetch, so the retention of fetched page text is not stated.
- **Legal basis:** Art. 6(1)(b) GDPR (contract performance); Art. 28 GDPR DPA (Staan's standard DPA, on file). The DPA states that search queries are not personal data and fall outside its scope (§2).
- **Third countries:** Staan's FAQ says an unnamed US fallback can supplement results, so some queries may leave the EU; the DPA does not list it. CIATA has asked Staan to switch it off for its account (issue #194).
- **DPA:** On file (Staan DPA, `signed_dpas/staan/`). Staan's standard DPA from its account console, no signature step; whether accepting the Terms of Service at registration concluded it is for counsel to confirm.

⚠ **OPEN:** The DPA does not cover the `full_content` page fetch, and gives 25 months for queries. Ask counsel whether derived queries can be personal data for CIATA (for example when a research question names a person), in which case Staan's §2 position does not hold.

⚠ **TO CONFIRM:** Confirm that the production environment has `SEARCH_BACKEND=staan` set in the production `.env` or `config.yaml` override. The code default in `SearchConfig` is "staan", but this should be verified against the actual running configuration. If a non-EU backend is active, the EU-only transfer claim does not hold for search processing.

5.3 Hetzner Online GmbH (Infrastructure)

- **Location:** Germany (Hetzner **Cloud** — shared virtualised infrastructure, not dedicated hardware)
- **Role:** Server hosting and storage — no access to application-layer data content.
- **DPA:** Required — not yet executed (Hetzner standard AVV / order-processing agreement).

✓ **CONFIRMED (2026-07-15):** The Hetzner Cloud production volumes are **not** encrypted at rest (no default volume/root-disk encryption). This is a documented, accepted gap. Compensating controls: bcrypt-hashed passwords, TLS in transit, strict per-user application-level access isolation, automated 30-day retention cleanup, and **encrypted off-site backups stored geo-redundantly within the EU (Hetzner Frankfurt → Helsinki)**. Enabling volume-level (LUKS/dm-crypt) at-rest encryption before the hot-production cutover remains a recommended hardening step.

5.4 Mollie B.V. (Card Payments)

- **Location:** Amsterdam, Netherlands (EU)
- **Data sent:** Payment amount, currency, description, metadata (`user_id`, `pack_id`, `credits`, `locale`), and — for Billie B2B method only — the buyer's billing address.

- **Retention:** Mollie retains payment records for its own regulatory, fraud-prevention, and financial-reporting purposes, independently of CIATA's retention schedule.
- **Legal basis:** Art. 6(1)(b) GDPR (contract performance); Art. 6(1)(c) (Mollie's legal obligations under financial regulation).
- **DPA:** Mollie standard DPA — required, not yet executed.

5.5 Billie GmbH (B2B Invoice / Net-30 BNPL — embedded in Mollie checkout)

- **Location:** Berlin, Germany
- **Trigger:** Activated only when the buyer selects the Billie payment method in the Mollie checkout. Not involved in standard card purchases.
- **Data sent:** Full billing address including given name, family name, organisation name, street, city, postal code, country, Handelsregister registration number, and VAT number — required for Billie's credit underwriting process.
- **Retention:** Billie retains buyer data for its **own** underwriting, credit risk assessment, anti-fraud, and regulatory purposes, independently of and beyond CIATA's retention schedule.
- **Legal basis:** Art. 6(1)(b) GDPR (Billie's contract with the buyer); Art. 6(1)(c) (Billie's legal obligations). Billie likely acts as an independent controller for underwriting data rather than as CIATA's processor.
- **DPA / controller relationship:** Billie acts as an independent controller for its BNPL credit decision (reached via Mollie's checkout, not a direct CIATA integration); on that basis no Art. 28 DPA is required. CIATA's obligation is Art. 13/14 disclosure, provided in the privacy policy (§6).

⚠ **TO CONFIRM (classification):** Confirm with counsel that Billie is correctly treated as an independent controller (BNPL, reached via Mollie). On that basis no Art. 28 DPA is required; Art. 13/14 disclosure is provided in the privacy policy. If Billie is later deemed a processor for any flow, obtain the corresponding DPA.

5.6 Qonto SAS (Invoicing)

- **Location:** Paris, France (EU)
- **Trigger:** Called for **every** completed purchase to create a VAT-compliant invoice (see `src/api/qonto_invoicing.py`).
- **Data sent:** Buyer name, email address, full billing address (where provided), VAT number, Mollie payment ID. Used to create or match a Qonto client record and generate a finalised invoice.
- **Retention:** Qonto retains invoicing data for its **own** accounting and tax-compliance purposes, independently of CIATA's retention schedule. French commercial law requires accounting records for a minimum of 10 years.
- **Legal basis:** Art. 6(1)(b) GDPR (contractual obligation to issue an invoice under §§ 14, 14a UStG and French equivalent); Art. 6(1)(c) (legal obligation).

- **DPA / controller relationship:** Qonto is a licensed payment institution acting in its own name and is treated as an independent controller for the invoice records it issues; on that basis no Art. 28 DPA is required. CIATA's obligation is Art. 13/14 disclosure, provided in the privacy policy (§6).

△ **TO CONFIRM (classification):** Confirm with counsel that Qonto is correctly treated as an independent controller. On that basis no Art. 28 DPA is required; Art. 13/14 disclosure is provided in the privacy policy (§6). Optionally obtain written confirmation of Qonto's role via dpo@qonto.com.

5.7 Scaleway SAS (Transactional Email)

- **Location:** Paris, France (EU). Data hosted and processed exclusively in the **fr-par** (Paris) region; Scaleway TEM is not offered in any region outside the EU, and Scaleway states that no transfers outside the EU occur.
- **Legal entity:** Scaleway SAS, 8 rue de la Ville l'Evêque, 75008 Paris. RCS Paris 433 115 904. Subsidiary of the French Iliad Group (approx. 96%).
- **Active since:** the provider switch on 2026-09-26 (issue #152; the code has been in production since commit d1f9449, the switch is the `EMAIL_PROVIDER=scaleway` setting). Brevo SAS served as the email processor until this date and remains in the code as a rollback option (`EMAIL_PROVIDER=brevo`); after the switch no personal data is sent to Brevo. Planned wind-down: once the switch has run stably, close the Brevo account and request deletion of its logs and contacts; record the confirmation. Brevo's Art. 28 DPA was never executed.
- **Data sent:**
 - Purchase confirmation: recipient email address, buyer name, purchase details (credits purchased, pack label, price, new credit balance, Mollie payment ID, purchase date).
 - Report-ready mail (live since 2026-09-26, issue #159): recipient email address, name if given, report title, a short summary of the report (the first paragraph of its executive summary, up to approximately 500 characters), link to the report, and the deletion date of the report. This mail carries research content (title and summary); professional-secrecy holders (e.g. law firms under §203 StGB) should consider turning it off. The mail is on by default and can be disabled in the account settings under Notifications. It is sent only for completed reports of verified, active accounts.
 - Account verification and password reset: recipient email address, name if given, verification or reset link.
- **Retention:** Scaleway keeps sending logs for its own service operation. The retention period is governed by Scaleway's DPA and documentation.
- **Legal basis:** Art. 6(1)(b) GDPR (contract performance: delivering a purchase receipt, a verification mail, or a report-ready notification).
- **DPA:** Scaleway publishes a standard Data Processing Agreement, "Version of June 1st, 2024", covering all Scaleway services including TEM. URL: https://www-uploads.scaleway.com/DPA_2024_ENG_b0abb5cc26.pdf. Status for CIATA: pending, to be accepted and filed; not yet on file.

- **Sub-processors:** Scaleway maintains a sub-processor list at <https://www.scaleway.com/en/subprocessorlist/>. It includes data-centre colocation operators headquartered in the US (Equinix EMEA, Iron Mountain, Digital Realty), each listed with EU processing locations only. Whether a US-headquartered colocation operator could be compelled to disclose data is a question for counsel.
- **No link tracking:** Scaleway TEM has no link-tracking feature; links in outbound mails are delivered exactly as written.

△ **TO CONFIRM:** Retention period of Scaleway TEM sending logs (governed by Scaleway's DPA; do not state a number of days until confirmed).

△ **TO CONFIRM:** Scaleway DPA "Version of June 1st, 2024" to be accepted and filed with CIATA's sub-processor records.

5.8 Plausible Analytics — NOT a third-party sub-processor

- **Location:** CIATA-controlled infrastructure at trc.ciata.io (Hetzner Cloud DE) — **self-hosted Plausible instance**
- **Data sent / collected:** Page-view counts, referrer statistics, and click-level conversion events via Plausible's tagged-events extension (`script.tagged-events.js`). Conversion events fire when users click designated interface elements; the following event names are configured in `frontend/index.html`: Home+Get+Started, Home+Pricing+CTA, Home+Final+CTA (landing-page call-to-action buttons), View+Create+Account (register-tab activation), Account+Created (registration form submit). Report-view events (issue #189): Report+Shared (share-link button), Report+Downloaded+Markdown (markdown download button), Report+Downloaded+PDF (PDF download button), Report+Copied+Button (copy-to-clipboard button), and Report Text Copied (native copy of selected report text, fired via the Plausible JS API from `frontend/js/app.js`). A further intended event, Start+Research (research submit), is declared with a non-standard attribute name (`data-plausible-event` rather than `plausible-event-name`) and we have not confirmed that it fires at all; it is listed here for completeness rather than as an assertion that this event is collected. Each event transmits only the event name and page URL — no form field values, no user identifiers, no device fingerprints are included. The Report Text Copied event sends only the event name, never the copied text. No IP addresses are stored by the self-hosted Plausible instance; no cookies are set; no cross-site tracking.
- **Personal data processed:** None. Event names and URLs are not personal data.
- **Legal basis:** Art. 6(1)(f) GDPR (legitimate interest in aggregate product and conversion analytics); no personal data processed.
- **Sub-processor status:** Self-hosted Plausible runs on CIATA's own infrastructure. It is **not** a third-party sub-processor under Art. 28 GDPR; no data is transferred to Plausible's own servers.

△ **TO CONFIRM (ePrivacy):** Confirm with counsel the ePrivacy / TDDDG position for cookie-free, identifier-free click event tracking via a self-hosted analytics instance. The prevailing view for cookieless analytics that collects no personal data is that no consent requirement

applies, but this should be confirmed against the applicable national implementation of the ePrivacy Directive.

5.9 OpenObserve operational telemetry — PLANNED, NOT YET ENABLED

- **Status:** Not active. The exporter in `src/perf.py` transmits only when `OPENOBSERVE_API_KEY`, `OPENOBSERVE_ORG` and `OPENOBSERVE_STREAM` are all set. They are not set, so **no telemetry currently leaves the application**. This entry is recorded in advance of enabling.
- **Purpose:** Service reliability and report-quality monitoring (latency, failure counts, synthesis-quality counters).
- **Data that would be sent:** Two record shapes. Performance records — function name, duration in milliseconds, success/failure status, and the exception message where a call failed. Counter records — a metric name drawn from a fixed set defined in source, a numeric value, and optionally the internal `job_id`, a locale code, a line count, or a short fixed reason code.
- **Data that would NOT be sent:** No research question, no report text, no scraped page content, no source URLs, no email addresses, no user identifiers, no IP addresses. Metric names ending `_url` are labels only; no URL accompanies them.
- **Personal data assessment:** The `job_id` is an opaque per-run identifier that CIATA can correlate to a user through its own database. On that basis these records are treated as **pseudonymous personal data**, not anonymous, pending counsel's confirmation.
- **Residual concern:** The performance record carries the raw exception string. Exception messages are not intended to contain user data, but this cannot be categorically excluded — an error text could quote a URL or a fragment of processed content. Scrubbing or dropping this field before enabling is under consideration.
- **Legal basis (intended):** Art. 6(1)(f) GDPR — legitimate interest in service reliability and quality.
- **Sub-processor status: Undecided, and it changes the classification.** Self-hosted on CIATA infrastructure (the code default points at `localhost:5080`) → not a third-party sub-processor, on the same reasoning as Plausible in §5.8. A hosted OpenObserve service → an Art. 28 processor requiring an executed DPA and, depending on region, a transfer assessment.
- **Retention at destination:** Undetermined — depends on the hosting decision and the instance's own retention configuration.

△ **BEFORE ENABLING:** (1) decide the hosting model; (2) if third-party, execute an Art. 28 DPA and complete a transfer assessment; (3) decide the treatment of the exception error field; (4) add the disclosure to `/privacy.html`; (5) add entries to the Art. 30 record (`ROPA.md`) and the sub-processor register (`DPA_REGISTER.md`).

6. Non-Default Backends and Cross-Border Transfers

The default and configured search backend is **Staan** (EU-sovereign index by Ecosia + Qwant, configured in `SearchConfig.backend = "staan"`). Two additional backends exist in `src/searcher.py` and can be enabled via the `SEARCH_BACKEND` configuration variable or `config.yaml`:

Backend	Jurisdiction	EEA transfer
Staan (default)	EU (European Perspective SAS)	No cross-border transfer
DuckDuckGo	US	EEA → US transfer; Art. 46 GDPR SCCs or equivalent required
Google Custom Search	US (Google LLC)	EEA → US transfer; Art. 46 GDPR SCCs or equivalent required

If DuckDuckGo or Google CSE is activated in production, the GDPR transfer basis must be documented and the privacy notice updated before processing begins. The technical safeguard "[x] EU-only data processing" in Section 7 is conditional on `SEARCH_BACKEND=staan` being the active production configuration.

⚠ TO CONFIRM: Verify the production `SEARCH_BACKEND` and `STAAN_FULL_CONTENT` settings in the live `.env` or `config.yaml` override to confirm **Staan** is the active backend and that no US-backend transfers are occurring without a documented transfer mechanism.

7. Technical Safeguards

- [x] PII filtering before LLM processing (email, phone, IBAN, credit card, DE/CH IDs, IPv4, DE postal addresses — see Section 4 for coverage limits)
- [x] No persistent storage of unfiltered raw page content — the PII filter (`scrub_pii`) runs before anything is written to disk: `page_text = scrub_pii(page_text).text` is executed at `src/engine.py:859` before `source.page_text = page_text`; raw text is never assigned to the field and therefore never reaches `state.json`
- [x] PII-filtered extractor input (up to 16,000 chars/source) is stored in `runs/{job_id}/state.json` for the 30-day run TTL; access is restricted to the authenticated user holding the admin role AND owning that specific run — the access-control check (`_check_job_ownership, src/api/app.py:279`) raises HTTP 404 for any other caller, including other admins; no staff-override path exists
- [x] Automated data retention with configurable TTL (default 30 days, `DATA_RETENTION_DAYS`)
- [x] GDPR Art. 17 deletion endpoint (`DELETE /v1/research/{job_id}`)
- [x] Ownership checks on all data access endpoints
- [x] TLS encryption in transit
- [x] Encrypted off-site backups, stored geo-redundantly within the EU (Hetzner Frankfurt → Helsinki)

- [] Disk encryption at rest — **not enabled** (confirmed 2026-07-15; see Section 5.3). Accepted gap with compensating controls (encrypted EU backups, per-user access isolation, 30-day retention); volume-level (LUKS) encryption is a recommended pre-launch hardening step
 - [x] No cookies or cross-site tracking; cookie-free self-hosted Plausible analytics (trc.ciata.io), no personal data
 - [x] No AI model training on user data (AKI.io DPA)
 - [x] EU AI Act Art. 50(2) transparency: (a) visible "AI-generated" label and provenance HTML comment in delivered reports; (b) invisible VSRMark Unicode variation-selector watermark embedded in report text (src/watermark.py). C2PA content provenance standard is **planned but not yet implemented** — it must not be presented as an existing measure.
 - [x] EU-only data processing (no cross-border transfers) — **conditional on** SEARCH_BACKEND=staan **in production** (see Section 6)
 - [x] Fonts and third-party JavaScript (marked.js) self-hosted since 2026-07-10; Google Fonts CDN and jsDelivr CDN removed
 - [x] robots.txt compliance checked before every direct page fetch; per-domain TTL cache (1 hour); graceful-allow fallback if robots.txt is unreachable (timeout or 404)
 - [x] Per-domain rate limiting (minimum 1 second between direct fetches to the same domain)
 - [x] Honest bot User-Agent declared in all direct page requests (CIATA-DeepResearch/1.0)
-

8. Retention Enforcement

Automated Cleanup

- **Mechanism:** Background timer runs every 24 hours (threading.Timer(86400, ...) in src/api/job_store.py)
- **Retention clock:** The created_at timestamp written to job_meta.json at run start governs expiry; if job_meta.json is absent, created_at in state.json is used; filesystem mtime is the fallback for legacy runs predating either field. See _get_run_created_at() in src/api/job_store.py.
- **Configuration:** DATA_RETENTION_DAYS environment variable (default: 30)
- **Scope:** Deletes entire run directory (state.json, report.md, results.tsv)
- **Verification:** Cleanup events logged at INFO level

On-Demand Deletion

- **User-initiated:** DELETE /v1/research/{job_id} API endpoint
- **Authorization:** Requires authentication + ownership verification
- **Scope:** Immediate deletion of run directory and in-memory records
- **Response:** HTTP 204 No Content on success

Account Deletion

- **Contact:** privacy@ciata.io
 - **Scope:** All account data and all associated research runs
 - **Timeline:** Within 30 days of request
 - **Note:** Payment records held by Mollie, Qonto, and Billie for their own regulatory purposes are not deleted upon CIATA account deletion; data subjects must contact those processors directly for erasure requests that fall within their independent obligations.
-

9. Audit Evidence

To verify compliance:

```
# Check current retention setting
echo $DATA_RETENTION_DAYS # Default: 30

# Show the retention clock (created_at) for each run.
# job_meta.json governs; state.json created_at is the secondary source;
# filesystem mtime is the fallback for legacy runs predating either field.
# This matches the logic in _get_run_created_at() (src/api/job_store.py).
python3 -c "
import json, pathlib
from datetime import datetime, timezone
runs = pathlib.Path('runs')
for sf in sorted(runs.glob('*state.json'))[:20]:
    job_id = sf.parent.name
    meta = sf.parent / 'job_meta.json'
    src = 'filesystem-mtime (legacy)'
    ts = datetime.fromtimestamp(sf.stat().st_mtime, tz=timezone.utc)
    if meta.exists():
        try:
            m = json.loads(meta.read_text())
            if m.get('created_at'):
                ts = datetime.fromisoformat(m['created_at'])
                if ts.tzinfo is None:
                    ts = ts.replace(tzinfo=timezone.utc)
                src = 'job_meta.json'
        except Exception:
            pass
    print(f'{job_id}: {ts.isoformat()} (source: {src})')
"
```

```
# Verify cleanup logs
docker logs auto_deep_research-api-1 | grep "Data retention"
```

```
# Report stored PII-filtered extractor input across all runs.
# Non-zero counts confirm the stored text is the PII-filtered extractor input;
# zero counts indicate legacy runs predating the feature.
# Unfiltered raw page content is never written to disk.
python3 -c "
import json, pathlib
```

```

runs = pathlib.Path('runs')
total_sources = 0
with_stored_text = 0
for sf in runs.glob('*/state.json'):
    try:
        data = json.loads(sf.read_text(encoding='utf-8'))
        for s in data.get('sources', []):
            total_sources += 1
            if s.get('page_text'):
                with_stored_text += 1
    except Exception:
        pass
print(f'{with_stored_text}/{total_sources} sources carry stored PII-filtered extractor input')
"

# Verify active search backend in config
grep -E "SEARCH_BACKEND|backend" .env config.yaml 2>/dev/null

# Verify Staan full_content setting
grep -E "STAAN_FULL_CONTENT|staan_full_content" .env config.yaml 2>/dev/null

# Verify VSRMark watermark in a report
python -m src.watermark verify runs/<job_id>/report.md

```

⚠ **TO CONFIRM – DPO Designation:** Assess whether CIATA GmbH is required to designate a Data Protection Officer under Art. 37 GDPR (criteria: core activities involve large-scale processing of special categories of data, or large-scale systematic monitoring of individuals). Confirm with legal counsel. If not required, document the assessment.

⚠ **TO CONFIRM – External Legal Review:** This statement has not been reviewed by external legal counsel. Review by a qualified data-protection lawyer is recommended before this document is published, shared with customers, or cited in controller records of processing activities (Art. 30 GDPR).

Approved by: Prof. Dr. Tim Landgraf, Managing Director, CIATA GmbH **Date:** 2026-08-11